Gradient Descent With Momentum.

velocity

$V_{dw} = \beta V_{dw} + (1-\beta)\, dw$ acceleration

$V_{db} = \beta V_{db} + (1-\beta)\, db$

acceleration (next gradient)

changes current velocity (momentum) by some extent.

$W = W - \alpha V_{dw}$, $b = b - \alpha V_{db}$

Hyperparameters : α , β $\beta = 0.9$ (average our last 10 gradients)

RMSprop

$S_{dw} = \beta_2 S_{dw} + (1-\beta_2)\, dw^2$

$S_{db} = \beta_2 S_{db} + (1-\beta_2)\, db^2$

$W := W - \alpha \frac{dw}{\sqrt{S_{dw}+\varepsilon}}$

$b := b - \alpha \frac{dw}{\sqrt{S_{db}+\varepsilon}}$

Adam optimization algorithm.

$V_{dw} = 0$, $S_{dw} = 0$, $V_{db} = 0$, $S_{db} = 0$

On iteration t :

Compute dw, db through mini-batch.

$V_{dw} = \beta_1 V_{dw} + (1-\beta_1)\, dw$, $V_{db} = \beta_1 V_{db} + (1-\beta_1)\, db$ ← "momentum" β_1

$S_{dw} = \beta_2 S_{dw} + (1-\beta_2)\, dw^2$, $S_{db} = \beta_2 S_{db} + (1-\beta_2)\, db^2$ ← "RMSprop" β_2

$V_{dw}^{corrected} = V_{dw} / (1-\beta_1)^t$, $V_{db}^{corrected} = \ldots$

$S_{dw}^{corrected} = \ldots$, $S_{db}^{corrected} = \ldots$

$$W := W - \alpha \frac{V_{dw}^{corrected}}{\sqrt{S_{dw}^{corrected}} + \varepsilon}$$

$b := \ldots$

Hyperparameter choice :

→ α : needs to be tune

β_1 : 0.9 → (dW)

β_2 : 0.999 → (dW^2)

ε : 10^{-8} Avoid divided by zero Exception.